# Change Point Detection in Trend of Mortality Data

F. Amani[a] , A.*Kazemnejad*[b] , *R.Habibi*[c]

[a] Department of Biostatistics, Ardabil University of Medical Science, Ardabil, Iran
[b] Department of Biostatistics, Tarbiat Modares University, Tehran, Iran
[c] Department of Statistics, Central Bank of Iran, Tehran, Iran

**Abstract** - Mortality refers to death that occurs within a population. It is linked to many factors such as age, sex, race, occupation and social class. Change in the pattern of mortality trend can affect the population standards of living and health care. This event makes a change point is occurred in mortality rates. The aim of this study is to detect change point in Iranian mortality data during1970 to 2007. We use several frequencies and Bayesian methods to estimate the change point in mortality data by Poisson model. All methods show that a change has been occurred in mortality rates of Iran at year 1993. Results showed that pattern and rate of mortality before and after change point is not similar.

**Keywords:** change point, mortality, Poisson, *Bayes information* criterion

## 1. Introduction

The change point problem arises in various practical fields such as Epidemiology, Toxicology, Medical, Economical surveys, Quality control, Statistical process control, Natural events, Demography and Mortality. Demography is statistical study of populations and vital statistics such as birth, death, marriage, divorce are important indicators in demography, so increase or decrease the rate of vital statistics due to change pattern of population [1]. Mortality, the death occurs within a population, is affected by several factors such as age, sex, race, occupation and social class. Therefore, change in population pattern due to difference distribution of mortality in one period compare to another period.

Let $\boldsymbol{X}_1,...,\boldsymbol{X}_{\boldsymbol{k}_0},\boldsymbol{X}_{\boldsymbol{k}_0+1},...,\boldsymbol{X}_{\boldsymbol{n}}$ be a sequence of independent observations which $\boldsymbol{X}_1,...,\boldsymbol{X}_{\boldsymbol{k}_0}$ have the Poisson distribution with rate $\lambda_0$ and remaining observations $\boldsymbol{X}_{k_0+1},...,\boldsymbol{X}_{\boldsymbol{n}}$ have the Poisson distribution with parameter $\lambda_1$ ($\lambda_0 \neq \lambda_1$), that is

$$X_i \sim \begin{cases} poi(\lambda_0) & i=1,...,k_0, \\ poi(\lambda_1) & i=k_0+1,...,n. \end{cases}$$

This problem is referred as the change point detection in statistical literatures. We test the null hypothesis of no change point $H_0:\lambda_0=\lambda_1$ against $H_1:\lambda_0\neq\lambda_1$ which stands a change point has been occurred in unknown point $k_0=1,...,n-1.$ If $H_0$ is rejected, then we estimate $\lambda_0,\lambda_1,k_0$.

During the last four decades, change point analysis has been received considerable attentions from both theoretical and practical aspects. Page proposed the change point problem in the context of quality control [2,3]. Hinckley [4] and Pettit [5] recommended the application of cumulative sum (CUSUM) to detect change point. The problem change in the mean of normal variables have been studied by Chernoff and Zacks [6], Sen and Srivastava [7], Hawkins [8] , Worsley [9] and Yao and Davis [10]. Some literatures about change point detection in Poisson distribution random variables were Raftery and Akman[11] ; Loader [12] ; Akman and Raftery [13] .

In this paper, change point analysis is done on Iranian mortality rates from 1971 to 2007 under by Poisson model. We first test the existence of change point and then estimate the location of change point.

## 2. Methodology

### A. Data Source and Collection Methods

Vital statistics such as birth, death, marriage, divorce and migration are the best direct indexes to estimate population changes in more countries. These data have been registered and collected by National Organization for Civil Registration (NOCR) in Iran and annually published by Statistical Center of Iran (SCI), a national government organization [14]. Another direct procedure for recording information about the members of a given population is census that conducted every ten years by SCI in Iran. The census data is commonly used for research, business marketing, planning, sampling surveys, demographic indexes and assessment vital statistics.

### B. Methods for change point detection in Poisson model

#### B.1. Cumulative Sum

A useful tool for change point detection in average of observations is the cumulative sum (CUSUM) test defined by

$$S_k = \sum_{i=1}^{k} (X_i - \bar{X}) / \mathbf{s}, \quad k = 1,...,n-1,$$

where $\bar{X}$ and **s** are the mean and standard deviation of all observations.

If there is no change point, cumulative sum should be small, and the null hypothesis is rejected for large value of CUSUM [2]. When the CUSUM is plotted against time, firstly we detect certain thresholds and when the value of $S_k$ exceeds a certain threshold value, we say that a change has been occurred in data trends. The CUSUM test statistic is given by

$$T = \max_{1 \le k \le n-1} |S_k|$$

If $H_0$ (there is no change point) is rejected, the location of change point is estimated by $S_{\hat{k}}$ that is the maximum value of $S_k$.

**B.2. Likelihood ratio (LR)**

One of the widely used methods in statistical hypothesis testing is LR. Here, the likelihood function under $H_0$ is

$$L_0(\lambda_0) = (1/\prod_{i=1}^{n} x_i !) e^{-n\lambda_0} \lambda_0^{\sum_{i=1}^{n} x_i}$$

and the MLE of $\lambda_0$ is given by $\hat{\lambda}_0 = (1/n)\sum_{i=1}^{n} x_i$ .

The likelihood function under $H_1$ is

$$L_1(\lambda_0, \lambda_1, k_0) = \frac{1}{\prod_{i=1}^{n} x_i !} \mathrm{e}^{-\{\lambda_0 k_0 + \lambda_1 (n-k_0)\}} \lambda_0^{\sum_{i=1}^{k_0} x_i} \lambda_1^{\sum_{i=k_0+1}^{n} x_i},$$

and for fixed $k_0$, the MLE's of $\lambda_0, \lambda_1$ are given by

$\hat{\lambda}_0 = (1/k_0)\sum_{i=1}^{k_0} x_i$ and $\hat{\lambda}_1 = (1/(n-k_0)) \sum_{i=k_0+1}^{n} x_i$ .

Then, the test statistic is given by

$$\boldsymbol{L} = \max_{1 \le \boldsymbol{k}_0 \le \boldsymbol{n}-1} \boldsymbol{L}_{\boldsymbol{k}_0},$$

where

$$L_{k_0} = -2 \log \frac{L_0(\hat{\lambda}_0)}{L_1(\hat{\lambda}_0, \hat{\lambda}_1)} .$$

The null hypothesis is rejected if $L > C$, where C is some constant which will be determined by the null distribution of $L$. Usually, the null distribution of L is too complicated and it is simulated by re sampling methods such as Monte Carlo technique or bootstrap method. If $H_0$ is rejected, then the position of change point is estimated by $\hat{k}$ where

$$L_{\hat{k}} = \max_{1 \le k \le n-1} L_k .$$

**B.3. Information criteria**

In statistics, information criteria (IC) are used in model selection. The IC can be applied to detect change point [15]. There are several types of IC which differ in their penalty terms. Akaike [16] proposed the Akaike information criterion (AIC). One of the modifications of AIC is the Bayesian information criterion that was proposed by Schwarz [17] and Chen and Gupta [15] and for simplify denoted by BIC. We used BIC to detect change point in Poisson model. The BIC under $H_0$ (no change point) and $H_1$ (there is a change point) are

$$BIC_{H_0} = -2 \log L_0(\hat{\lambda}_0) + 2 \log n ,$$

$$BIC_{H_1}(k_0) = -2 \log L_1(\hat{\lambda}_0, \hat{\lambda}_1) + 2 \log n .$$

We will reject $H_0$ if

$$BIC_{H_0} > \min_{1 \le k_0 \le n-1} BIC_{H_1}(k_0) .$$

The change point position estimated by $\hat{k}$ that

$$BIC_{H_1}(\hat{k}) = \min_{1 \le k_0 \le n-1} BIC_{H_1}(k_0) .$$

**B.4. Bayesian approaches**

The Bayesian method is one of the best methods used for detect change point in data set. The Bayesian approach assumes priors for parameters and uses them to obtain posterior. Inferences about the change point and the parameters before and after change point, using a Bayesian approach by assume prior distributions for the parameters as follows

$$\pi(\lambda_0) \sim \Gamma(a_1, b_1) , \pi(\lambda_1) \sim \Gamma(a_2, b_2) , \pi(k_0) \sim DU(1,...,n-1)$$

where $\Gamma$ is gamma , DU is discrete uniform and IG is inverse gamma distributions.

Therefore, the five dimensional posterior distributions is given by

$$f(k_0, \lambda_1, \lambda_2 | Y) \propto \prod_{i=1}^{k_0} \frac{\lambda_0^{Y_i} e^{-\lambda_0}}{Y_i !} \prod_{i=k_0+1}^{n} \frac{\lambda_1^{Y_i} e^{-\lambda_1}}{Y_i !} .$$

Markov Chain Monte Carlo (MCMC) is applied to generate samples from posterior distribution[18] to estimate necessary parameters such as $k_0, \lambda_1, \lambda_2$ that given by

$$f(\lambda_0 | k_0, \lambda_2, b_1, b_2, Y) \propto \Gamma(\sum_{i=1}^{k_0} Y_i + 0.5, \frac{b_1}{k_0 b_1 + 1})$$

$$f(\lambda_1 | k_0, \lambda_0, b_1, b_2, Y) \propto \Gamma(\sum_{i=k_0+1}^{n} Y_i + 0.5, \frac{b_2}{(n-k_0) b_2 + 1})$$

$$f(k_0 | \lambda_0, \lambda_1, b_1, b_2, Y) \propto \lambda_0^{\sum_{i=1}^{k_0} Y_i} \lambda_1^{\sum_{i=k_0+1}^{n} Y_i} e^{-k_0 \lambda_0 - (n-k_0)\lambda_1}$$

## 3. Results

### 3.1. Description of mortality data

A total of 9.7 million deaths registered in during 1971 to 2007 in Iran. About 5.9 million (61.1 %) were in male and 3.8 million (38.9%) in female with sex ratio 1.6 that from them 5.86 million (60.4%) were in urban and 3.84 million (% 39.6) in rural. Crude death rate in all studied population according to census statistics during 1971 to 2010 was 5.1 per 1000 people. The results showed that the CDR rate in Iran was 13 in years 1970-75 and dropped to 5 in 2005-2010 (see Figure 1). Life expectancy at birth increased during the third (57.7) and fourth (59.6) censuses, but during recent ten years, it was increased more rapidly and reached to 71 in years 2005-2010 (see Figure 2). This descriptive result suggests the existence of change point among the data. To check these possibilities, we first test the existence of change point and then we estimate it.

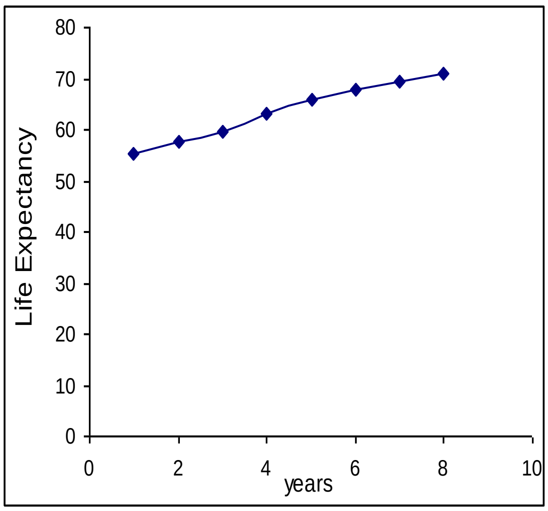


**Fig.1.** Life Expectancy in Iran 1970-2010

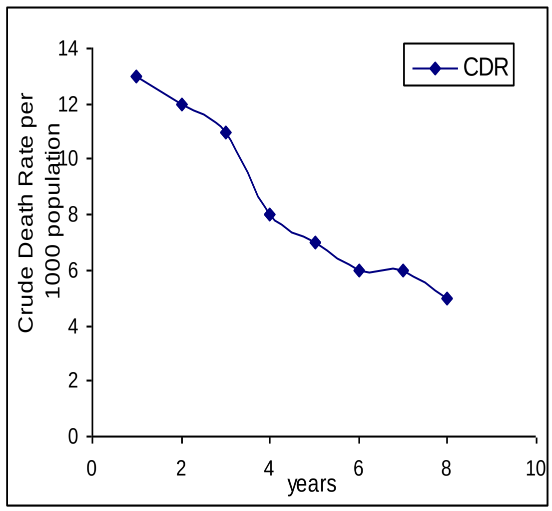


**Fig.2.** Crude Death Rate in Iran1970-2010

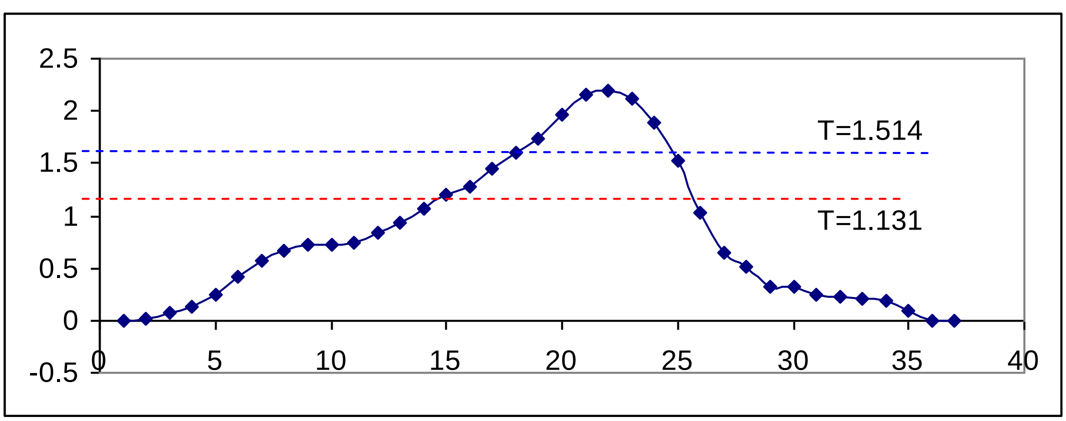


**Fig. 3.**Cusum chart for the mortality data

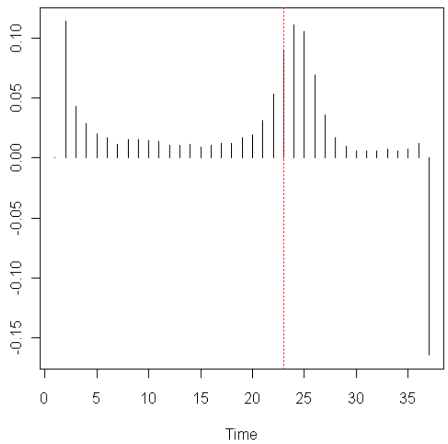


**Fig.4.** The posterior density of $k_0$

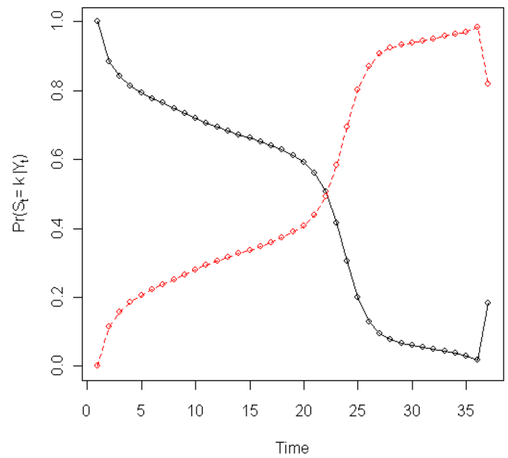


**Fig.5.** State of change point

**3.2. Change point in real data by CUSUM, LR and BIC in Poisson Distribution**

Here, we detect change point using CUSUM, LR and BIC methods. The CUSUM chart shows a change point has been occurred in $k_0 = 22$ i.e., in 1993 (see Figure 3). Three null quintiles of test statistic $n^{(-1/2)}T$ (with n=37 sample size of real data) in the levels of $\alpha = 0.1, 0.05$ and 0.01 are 1.131, 1.261 and 1.514, respectively. These values are obtained using simulation programs. Then, $H_0$ is rejected because the value of test statistics is 2.202 which is bigger than the above mentioned quintiles. For BIC the null hypothesis $H_0$ is rejected because

$$BIC_{H_0} = 225.1054 > \min_{1 \le k_0 \le n-1} BIC_{H_1}(k_0) = 216.78$$

.

For LR value, $H_0$ is rejected by comparing LR with its critical values. The estimated change point is $k_0 = 22$ with LR=-4.72 and parameters before and after change point $\lambda_0 = 4.27, \lambda_1 = 6.14$.

**3.3. Change point in real data using Bayes in Poisson model**

For Bayes method, we assign $\Gamma(1,1)$ priors for both Poisson means before and after change point. We estimated change at point $k_0 = 22$ and the values of rates before and after change are 4.08 and 5.36. Two Figures (4 and 5) are drawn using MCMC pack taken from R software. Figure 4 shows the posterior density of change point. It is seen that this density is maximized exactly at $k_0 = 22$ which says our estimate is Maximum A Posteriori (MAP), a good property for Bayesian estimators which is attained in our problem. Figure 5 is the plot of the posterior probability that each time point is in each state. The change point is too clear visually using this plot.

**3.4. Change point in simulated data using Poisson distribution and poison**

Here, using simulated data, we compare the performance of methods mentioned in previous sections. The procedures are compared by MSE of change point estimators. We assume the data before and after change point have Poisson distributions. Here, n=100. The results are presented in Table 1. Our simulation results show that the CUSUM procedure works well, although they are not given here.

**TABLE.1** Results of Simulation Data in Poisson distribution

| $k_0$ | $\lambda_0$ | $\lambda_1$ | $\hat{k}$ | $\hat{\lambda}_0$ | $\hat{\lambda}_1$ | $MSE(\frac{\hat{k}}{100})$ | $MSE(\hat{\lambda}_0)$ | $MSE(\hat{\lambda}_1)$ |
|---|---|---|---|---|---|---|---|---|
| 10 | 2 | 5 | 10 | 2.08 | 2.18 | 0.00000028 | 0.00022 | 0.0006 |
| 20 | 2 | 3 | 19 | 2.28 | 2.13 | 0.00016 | 0.0025 | 0.0009 |
| 30 | 2 | 2.1 | 31 | 2.04 | 2.08 | 0.000087 | 0.00023 | 0.00016 |
| 40 | 2 | 2.1 | 40 | 1.85 | 2.04 | 0.0000025 | 0.002 | 0.00034 |
| 50 | 2 | 2.1 | 49 | 2.04 | 2.08 | 0.0001 | 0.0016 | 0.0001 |
| 60 | 2 | 2.1 | 59 | 2.05 | 2.01 | 0.0003 | 0.008 | .006 |
| 70 | 2 | 2.1 | 69 | 2.01 | 2.09 | 0.001 | 0.008 | .000016 |
| 80 | 2 | 2.1 | 79 | 2.08 | 2.1 | 0.0009 | 0.00027 | .000006 |
| 90 | 2 | 2.1 | 88 | 2.07 | 2.06 | 0.0016 | 0.00013 | .0016 |

**4. Concluding remarks**

In this paper, we used frequentist and Bayesian methods to detect change point in Iranian mortality rates. With assume Poisson distribution results showed that a change has been occurred on mortality rates at $k_0 = 22$ in year 1993. This result corresponds to result of SCI, although their results are descriptive. All methods work well under Poisson model. Results showed that a change has been occurred in year 1993 at point $k_0 = 23$ and pattern of mortality data heve increasing trend from 1970 to 1993 with rate $\lambda_0 = 6.14$ and decreasing trend after 1993 with rate $\lambda_1 = 4.27$. Alos, we showed that in spite of existence many methods for analysis change in pattern of mortality data, statistical change point analysis is a useful tool for assessment mortality trend or all time related data such as vital statistics and demographic indicators.

**Acknowledgment**

We are tanks the personal of Iranian Statistical Center library personals for helping us for collected necessary real mortality data.

**Biographies:**

**Firouz amani** was born in west Azerbaijan, Iran in 1976. He was awarded the B.Sc., M.Sc. and Ph.D. degrees from the University of Kermanshah, Tehran, and Tehran in 1999, 2001 and 2010; respectively. In 2001, he joined the Ardabil University of Medical Science as a lecturer and was an Assistant Professor from 2010. His research interests are in Regression, change point, Demography and trend analysis. Address correspondence to dr.firouz.amani, Department of Biostatistics, Ardabil University of Medical Science, Ardabil, Iran. E-mail: f.amani@arums.ac.ir. 00984515510057-9144519673

**Anoushirvan Kazemnejad** was born in Tehran, Iran 1958. He was awarded the B.Sc, M.Sc. and Ph.D. degrees from the University of Tehran, Tehran and Tehran in 1986,1991,1996; respectively. In 1996, he joined the Tarbiat Modaress University as a lecturer and was an assistant professor from 2002. His research interests are in GLM, Survival analysis and clinical trial.

**Reza Habibi** works as Assistant Prof in Statistics, Department of Statistics, Central Bank of Iran, Tehran, Iran

.

**REFERENCES**

[1] Mohammad, I. (2006). Trends and patterns of mortality in china, Japan and India: A Comparative Analysis .The Social Science 1(2):149-153.

[2] Page, E.S. (1954). A test for a change in a parameter occurring at an unknown point. Biometrika 42: 523-527.

[3] Page, E.S. (1957). On problem in which a change in a parameter occurs at an unknown point.Biometrika 44:248-252.

[4] Hinkley, D.V. (1970). Inference about the change-point in a sequence of random variables. Biometrika 57(1): 1–17.

[5] Pettitt, A.N. (1979). A non-parametric approach to the change-point problem. Appl. Statist 28:126-135.
[6] Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal which is subjected to changes in time. Ann.Math.Statist 35: 999-1018.
[7] Sen, A.K. and Srivastava, M.S. (1975). On tests for detecting changes in mean, Ann. Statist 3: 98-108.
[8] Hawkins, D.M. (1977). Testing a sequence of observations for a shift in location. J.Amer. Statist.Assoc 72: 180-186.
[9] Worsley, K.J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. Biometrika 73: 91-104.
[10] Yao, Y.C. and Davis, R.A. (1984). The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. Sankhya Ser. A 48: 339-353.
[11] Raftery, A.E. and Akman, V.E. (1986). Bayesian analysis of a Poisson process with a change point. Biometrika 73: 85-89.
[12] Loader, C.R. (1992). A log-linear model for a Poisson process change point .The Annals of Statistics20: 1391-1411.
[13] Akman,V.E. and Raftery, A.E. (1986). Asymptotic inference for a change point Poisson process. The Annals of Statistics14:1583-1590.
[14] Khosravi, A., Taylor, R., Nagavi, M., and Lopez, A.D. (2007). Mortality in the Islamic Republic of Iran, 1964-2004. Bull World Health Organ 85:607-614.
[15] Chen, J. and Gupta, A.K. (2000). Parametric statistical change point analysis. Boston: Birkhauser.
[16] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6): 716-723.
[17] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6 (2): 461–464.
[18] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57 (1): 97–109.